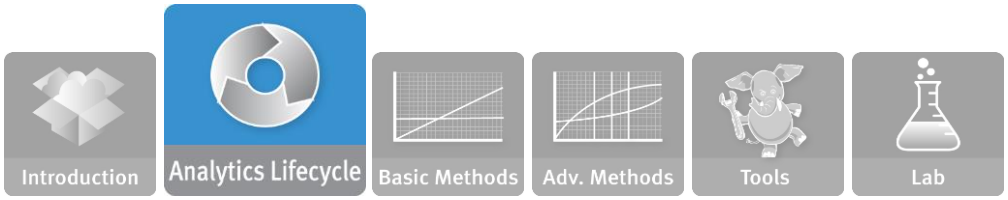
Information Systems

Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Module 2 – Data Analytics Lifecycle



Module 2: Data Analytics Lifecycle…..Continue

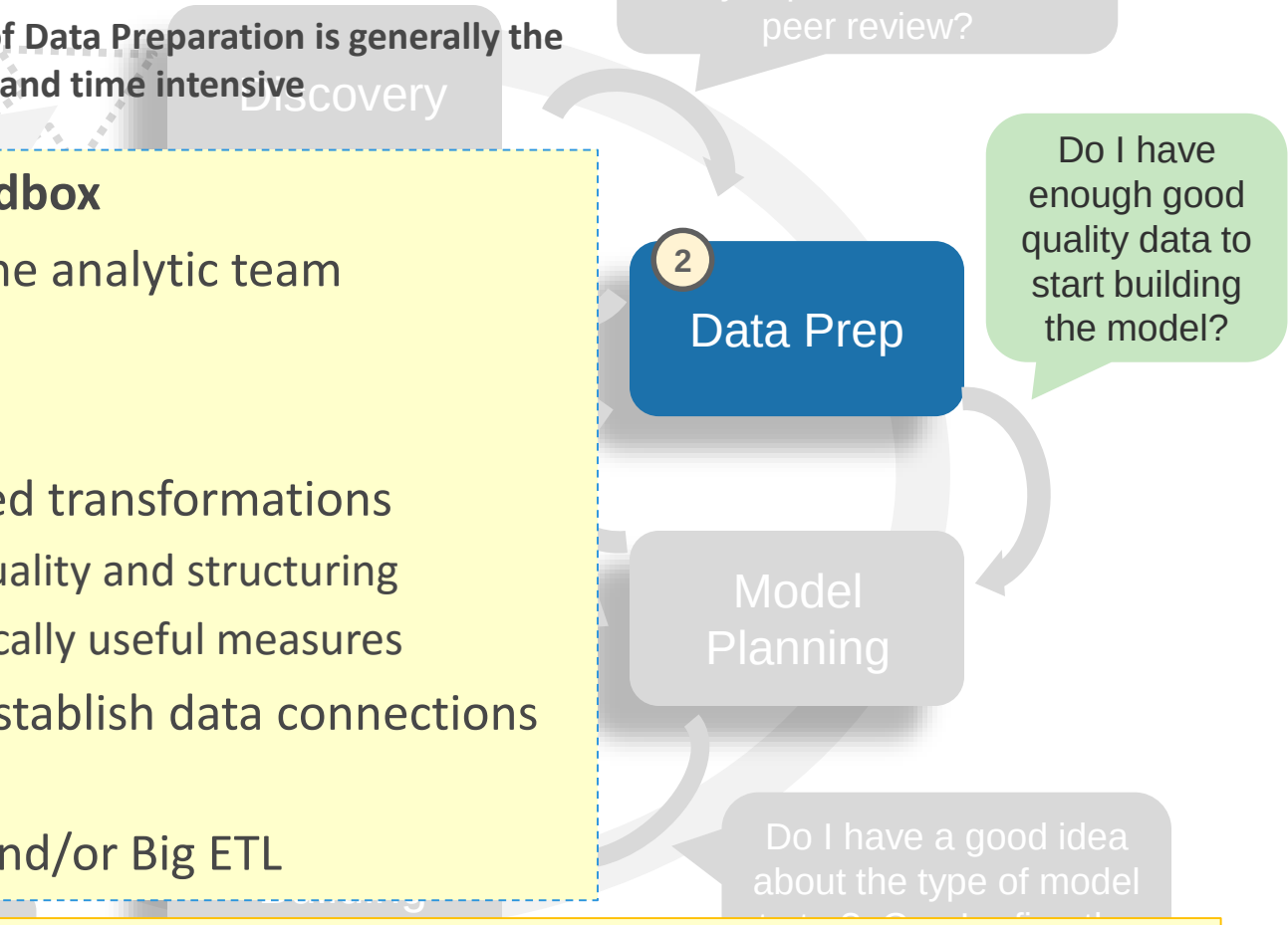
Upon completion of this module, you should be able to:

- Apply the Data Analytics Lifecycle to a case study scenario
- Frame a business problem as an analytics problem
- Identify the four main deliverables in an analytics project

Data Analytics Lifecycle
Phase 2: Data Preparation

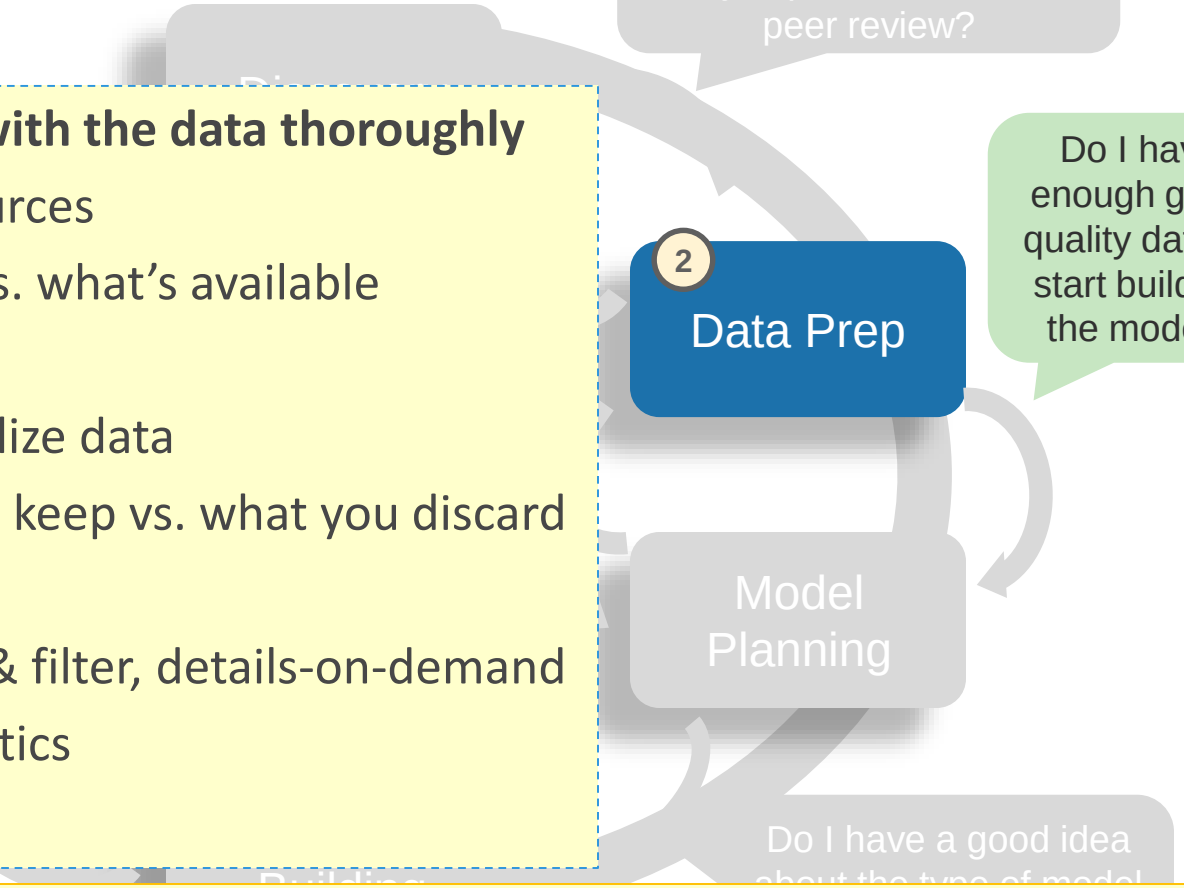
Of all of the phases, the step of Data Preparation is generally the most iterative and time intensive

- **Prepare Analytic Sandbox**
  - Work space for the analytic team
  - 10x+ vs. EDW
- **Perform ELT**
  - Determine needed transformations
    - Assess data quality and structuring
    - Derive statistically useful measures
  - Determine and establish data connections for raw data
  - Execute Big ELT and/or Big ETL



- **<u>Useful Tools for this phase:</u>**
  - ***For Data Transformation & Cleansing***: SQL, Hadoop, MapReduce, Alpine Miner

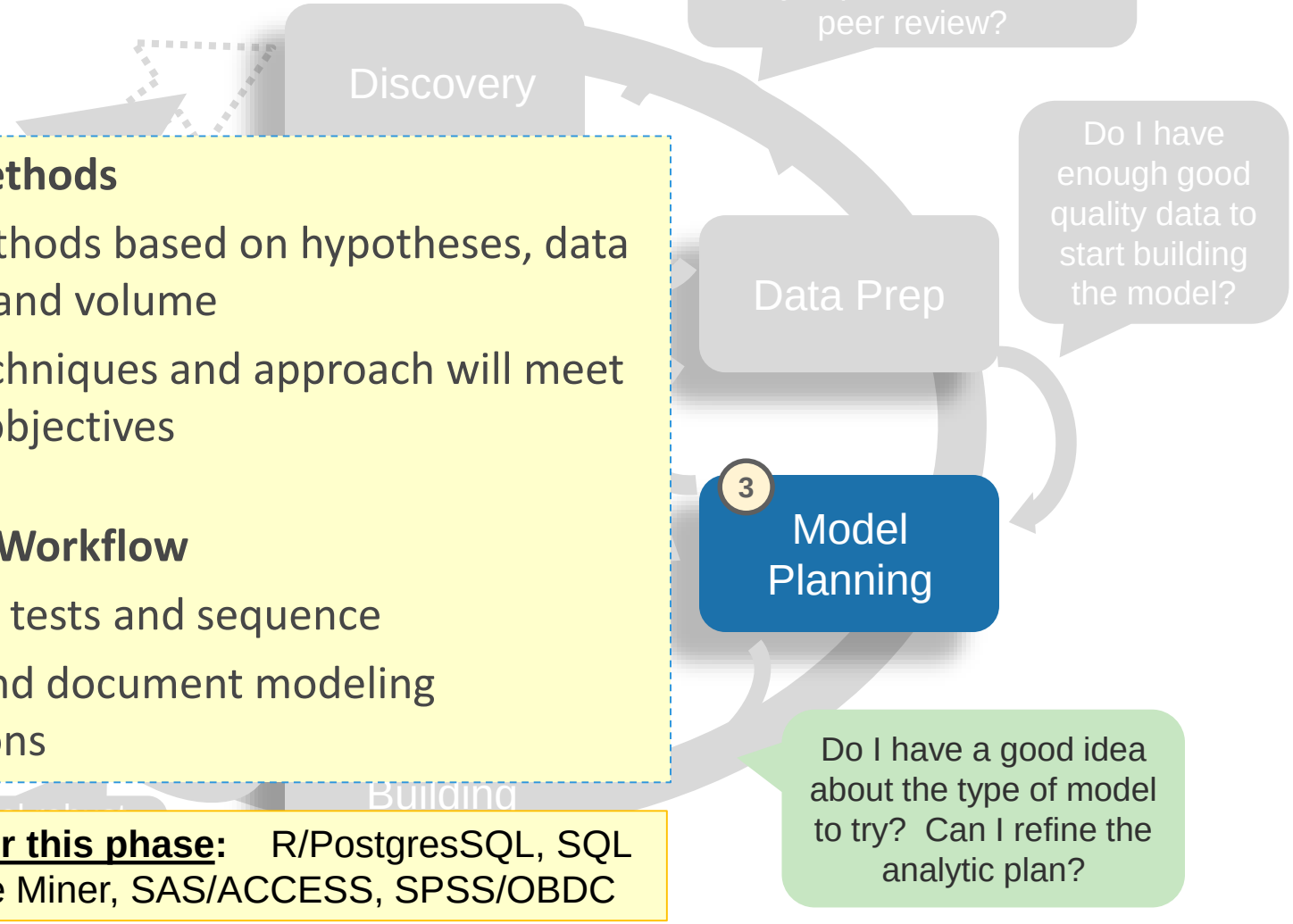
Data Analytics Lifecycle
Phase 2: Data Preparation



- **Familiarize yourself with the data thoroughly**
  - List your data sources
  - What's needed vs. what's available
- **Data Conditioning**
  - Clean and normalize data
  - Discern what you keep vs. what you discard
- **Survey & Visualize**
  - Overview, zoom & filter, details-on-demand
  - Descriptive Statistics
  - Data Quality

- **Useful Tools for this phase:**
  - Descriptive Statistics on candidate variables for diagnostics & quality
  - ***Visualization***: R (base package, ggplot and lattice), GnuPlot, Ggobi/Rggobi, Spotfire, Tableau

Data Analytics Lifecycle
Phase 3: Model Planning

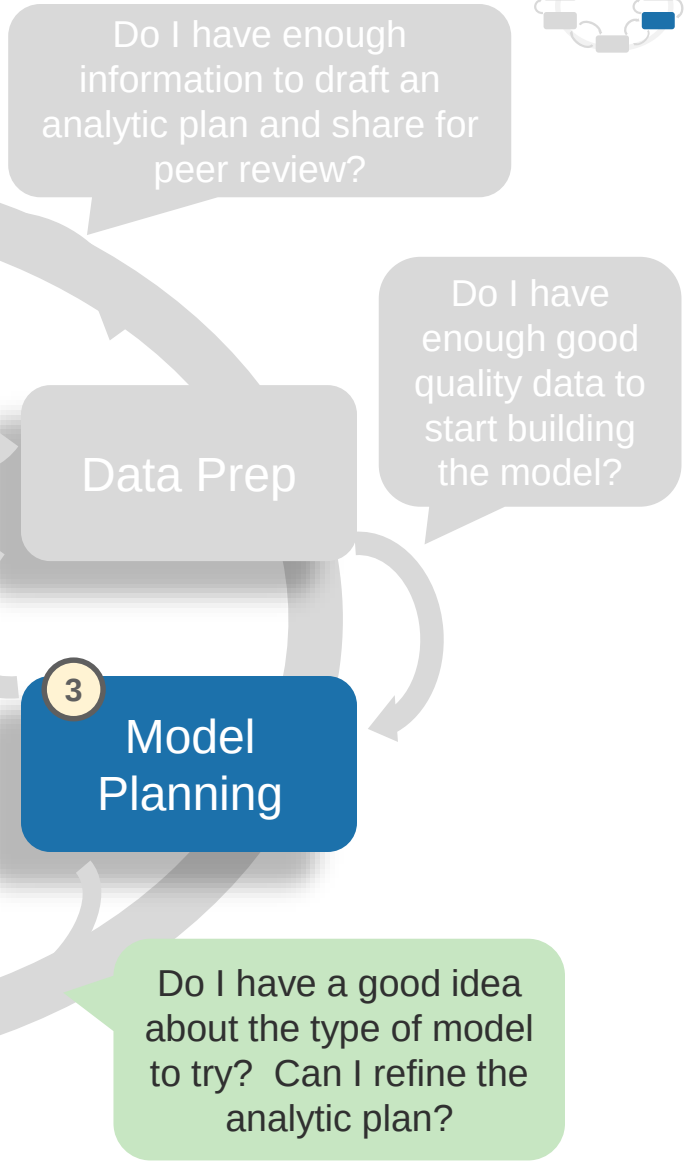


- **Determine Methods**
  - Select methods based on hypotheses, data structure and volume
  - Ensure techniques and approach will meet business objectives
- **Techniques & Workflow**
  - Candidate tests and sequence
  - Identify and document modeling assumptions

- **Useful Tools for this phase:** R/PostgresSQL, SQL Analytics, Alpine Miner, SAS/ACCESS, SPSS/OBDC

Data Analytics Lifecycle
Phase 3: Model Planning

- **Data Exploration**
- **Variable Selection**
  - Inputs from stakeholders and domain experts
  - Capture essence of the predictors, leverage a technique for dimensionality reduction
  - Iterative testing to confirm the most significant variables
- **Model Selection**
  - Conversion to SQL or database language for best performance
  - Choose technique based on the end goal and the data format



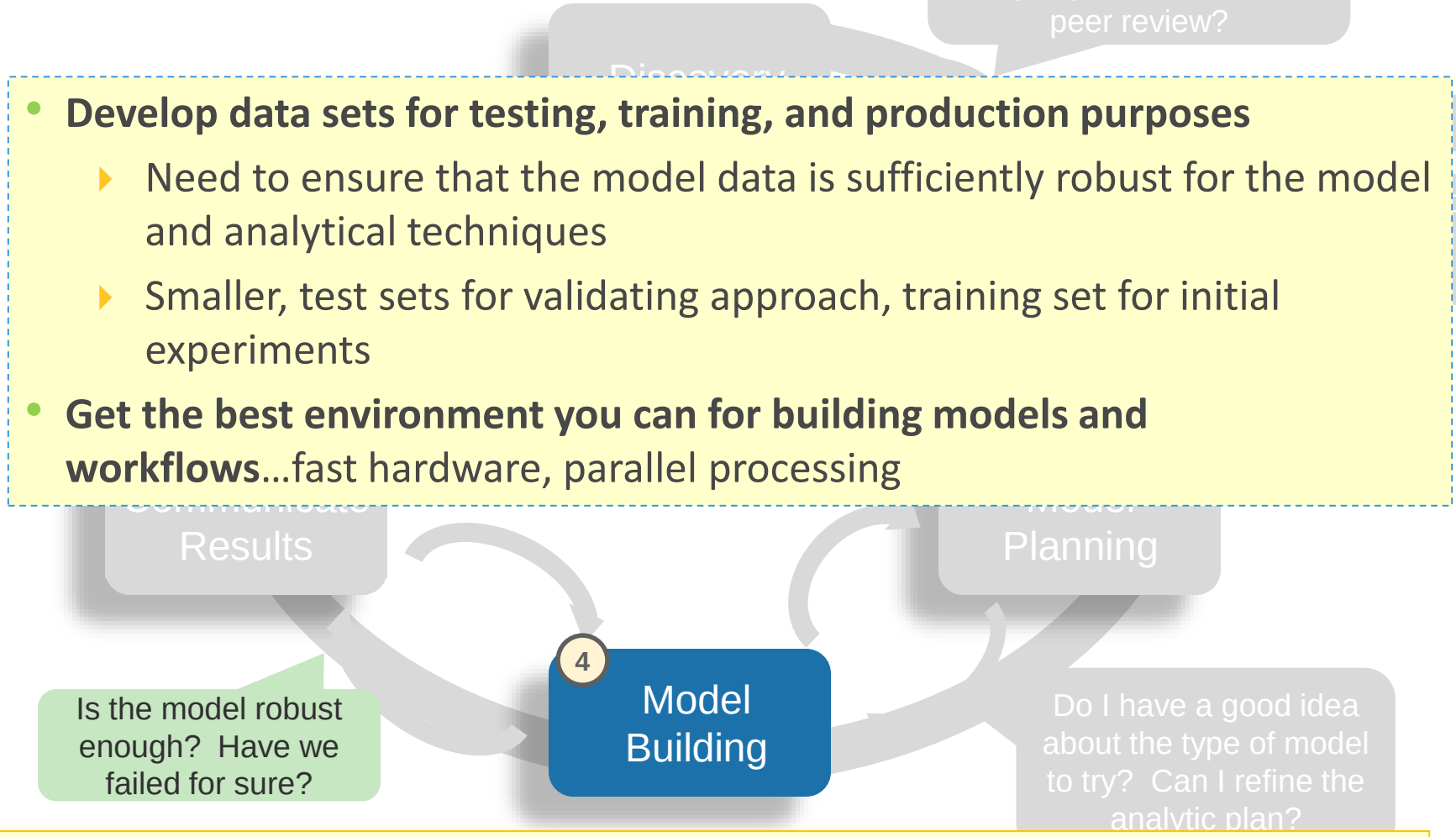
Sample Research: Churn Prediction in Other Verticals

Mini Case Study: Churn Prediction for Yoyodyne Bank

- After conducting research on churn prediction, you have identified many methods for analyzing customer churn across multiple verticals (those in **bold** are taught in this course)
- At this point, a Data Scientist would assess the methods and select the best model for the situation

Market Sector	Analytic Techniques/Methods Used
Wireless Telecom	DMEL method (data mining by evolutionary learning)
Retail Business	**Logistic regression**, ARD (automatic relevance determination), **decision tree**
Daily Grocery	MLR (**multiple linear regression**), ARD, and **decision tree**
Wireless Telecom	Neural network, **decision tree**, hierarchical neurofuzzy systems, rule evolver
Retail Banking	**Multiple regression**
Wireless Telecom	L**ogistic regression**, neural network, **decision tree**

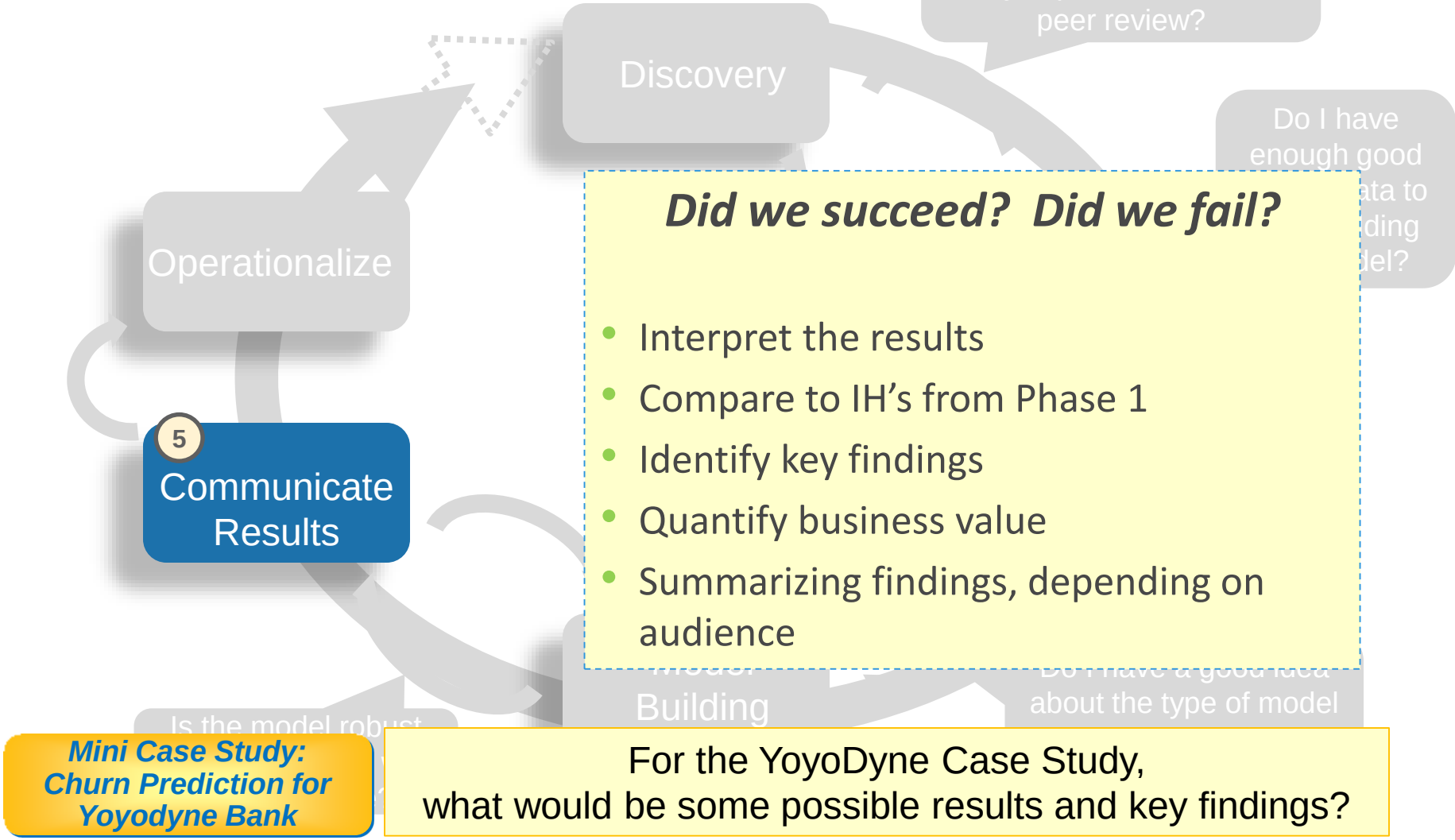
Data Analytics Lifecycle
Phase 4: Model Building



- **Develop data sets for testing, training, and production purposes**
  - Need to ensure that the model data is sufficiently robust for the model and analytical techniques
  - Smaller, test sets for validating approach, training set for initial experiments
- **Get the best environment you can for building models and workflows**…fast hardware, parallel processing

- **Useful Tools for this phase**: R, PL/R, SQL, Alpine Miner, SAS Enterprise Miner

Data Analytics Lifecycle
Phase 5: Communicate Results



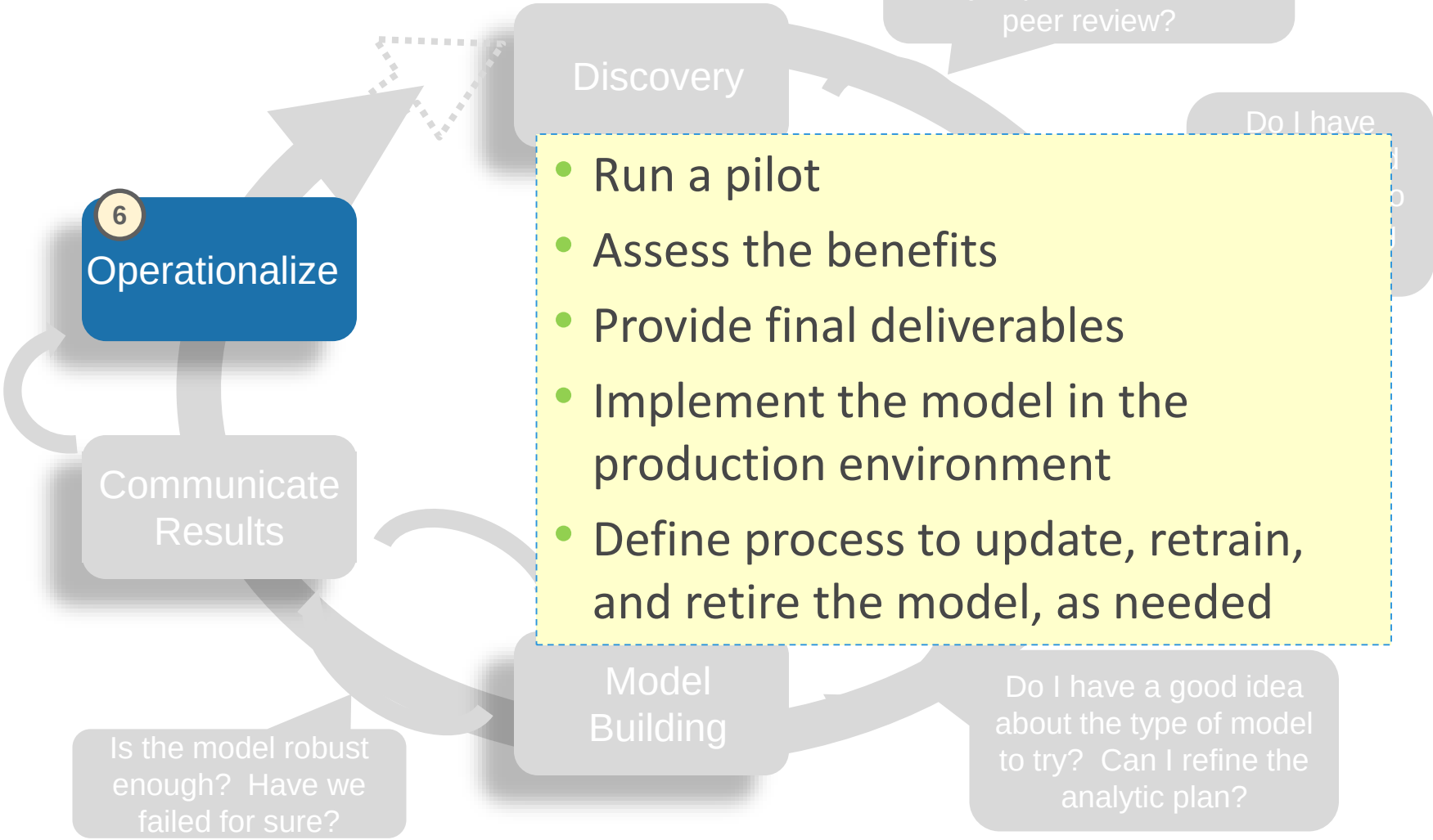
Did we succeed? Did we fail?

- Interpret the results
- Compare to IH's from Phase 1
- Identify key findings
- Quantify business value
- Summarizing findings, depending on audience

Mini Case Study: Churn Prediction for Yoyodyne Bank

For the YoyoDyne Case Study,
what would be some possible results and key findings?

Data Analytics Lifecycle
Phase 6: Operationalize



Analytic Plan

Mini Case Study: Churn Prediction for Retail Banking

Components of Analytic Plan	Retail Banking: Yoyodyne Bank
Phase 1: Discovery Business Problem Framed	How do we identify churn/no churn for a customer?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates.
Data	5 months of customer account history.
Phase 3: Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn.
Phase 5: Result & Key Findings	Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn. If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	If we can target customers who are high-risk for churn, we can reduce customer attrition by 25%. This would save $3 million in lost of customer revenue and avoid $1.5 million in new customer acquisition costs each year.

Key Outputs from a Successful Analytic Project, by Role

Role	Description	What the Role Needs in the Final Deliverables

Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized	• **Sponsor Presentation** addressing:
		  • Are the results good for me?
		  • What are the benefits of the findings?
		  • What are the implications of this for me?

Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team	• **Sponsor Presentation** addressing:
		  • What's the business impact of doing this?
		  • What are the risks? ROI?
		  • How can this be evangelized within the organization (and beyond)?

Project Manager	Ensure key milestones and objectives are met on time and at expected quality.	
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective	• Show the **analyst presentation** • Determine if the reports will change
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met	• Show the **analyst presentation** • Share the **code**

4 Core Deliverables to Meet Most Stakeholder Needs

1. **Presentation for Project Sponsors**
   - “Big picture" takeaways for executive level stakeholders
   - Determine key messages to aid their decision-making process
   - Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp
2. **Presentation for Analysts**
   - Business process changes
   - Reporting changes
   - Fellow Data Scientists will want the details and are comfortable with technical graphs (such as ROC curves, density plots, histograms)
3. **Code** for technical people
4. **Technical specs** of implementing the code

Analyst Wish List for a Successful Analytics Project

Data & Workspaces

- Access to all the data, including aggregated OLAP data, BI tools, raw data, structured and various states of unstructured data as needed
- Up-to-date data dictionary to describe the data
- Area for staging and production data sets
- Ability to move data back and forth between workspaces and staging areas
- Analytic sandbox with strong compute power to experiment and play with the data

Tools

- Statistical/mathematical/visual software of choice for a given situation and problem set, such as SAS, Matlab, R, java tools, Tableau, Spotfire
- Collaboration: an online platform or environment for collaboration and communicating with team members
- Tool or place to log errors with systems, environments or data sets

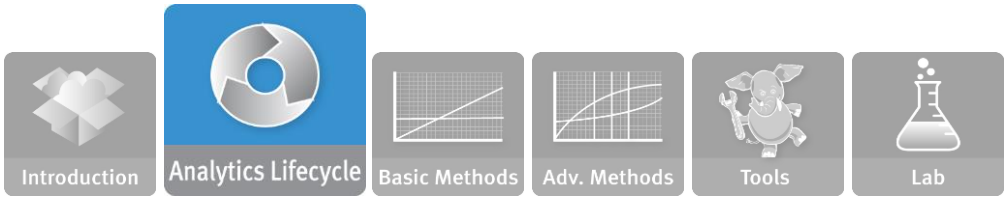
Check Your Knowledge

Your Thoughts?

- In which phase would you expect to invest most of your project time and why? Where would expect to spend the least time?

- What are the benefits of doing a pilot program before a full scale rollout of a new analytical methodology? Discuss this in the context of the mini case study.

- What kinds of tools would be used in the following phases, and for which kinds of use scenarios?
  - Phase 2: Data Preparation
  - Phase 4: Model Execution
- Now that you have completed the analytical project at Yoyodyne, you have an opportunity to repurpose this approach for an online eCommerce company. What phases of the lifecycle do you need to focus on to identify ways to do this?



Module 2: Summary

Key points covered in this module:

- The Data Analytics Lifecycle was applied to a case study scenario
- A business problem was framed as an analytics problem
- The four main deliverables in an analytics project were identified

Lab Exercise 1: Introduction to Data Environment

This first lab introduces the Analytics Lab Environment you will be working on throughout the course.

After completing the tasks in this lab you should be able to:

- Authenticate and access the Virtual Machine (VM) assigned to you for all of your lab exercises
- Locate data sets you will be working with for the course's labs
- Use meta commands and PSQL to navigate through the data sets
- Create sub-sets of the big data, using table joins and filters to analyze subsequent lab exercises

Thanks